"Variable selection in a flexible parametric mixture cure model with interval-censored data"

Scolas, Sylvie ; El Ghouch, Anouar ; Legrand, Catherine ; Oulhaj, Abderrahim

Abstract

In survival analysis, it is generally assumed that every individual will someday ex- perience the event of interest. However, this is not always the case, as some individuals may not be susceptible to this event. Also, in medical studies, it is frequent that pa- tients come to scheduled interviews and that the time to the event is only known to occur between two visits. That is, the data are interval-censored with a cure fraction. Variable selection in such a setting is of outstanding interest. Covariates impacting the survival are not necessarily the same as those impacting the probability to experience the event. The objective of this paper is to develop a parametric but flexible statistical model to analyze data that are interval-censored and include a fraction of cured indi- viduals when the number of potential covariates may be large. We use the parametric mixture cure model with an accelerated failure regression model for the survival, along with the extended generalized gamma ...

Document type : *Document de travail (Working Paper)*

I N S T I T U T D E S T A T I S T I Q U E
B I O S T A T I S T I Q U E E T
S C I E N C E S A C T U A R I E L L E S
(I S B A)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N
P A P E R

2014/41

Variable selection in a flexible parametric mixture cure model with interval-censored data

SCOLAS, S., EL GHOUCH, A., LEGRAND, C. AND A. OULHAJ

Variable selection in a flexible parametric mixture cure model with interval-censored data

Sylvie Scolas*[1], Anouar El Ghouch [1], Catherine Legrand [1], and Abderrahim Oulhaj [2]

[1]Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Université catholique de Louvain, Louvain-la-Neuve, Belgium
[2]College of Medicine & Health Sciences, United Arab Emirates University, Arab Emirates

Abstract

In survival analysis, it is generally assumed that every individual will someday experience the event of interest. However, this is not always the case, as some individuals may not be susceptible to this event. Also, in medical studies, it is frequent that patients come to scheduled interviews and that the time to the event is only known to occur between two visits. That is, the data are interval-censored with a cure fraction. Variable selection in such a setting is of outstanding interest. Covariates impacting the survival are not necessarily the same as those impacting the probability to experience the event. The objective of this paper is to develop a parametric but flexible statistical model to analyze data that are interval-censored and include a fraction of cured individuals when the number of potential covariates may be large. We use the parametric mixture cure model with an accelerated failure regression model for the survival, along with the extended generalized gamma for the error term. To overcome the issue of non-stable and non-continuous variable selection procedures, we extend the adaptive LASSO to our model. By means of simulation studies, we show the good performance of our method, and discuss the behavior of estimates with varying cure and censoring proportion. Lastly, our proposed method is illustrated with a real database studying the time until conversion to amnestic Mild Cognitive Impairment, a possible precursor of Alzheimer disease.

1 Introduction

Alzheimer's disease is one of the worst plagues of this century. Some factors, such as conversion to amnestic Mild Cognitive Impairment (aMCI), are nowadays considered as precursors

*sylvie.scolas@uclouvain.be

of the disease [1]. In the management of at risk population (i.e. elderly), it is therefore important to study the time to aMCI conversion, and to identify risk factors associated with it. Several studies were performed within this respect [2, 3, 4]. In particular, we consider here a study [5] conducted from 1988 to 2008 which included 241 healthy elderly people (average age of 72 years old) and presents several interesting features. Since participants were followed at regular interviews, the endpoint of interest in this study, the time to aMCI conversion, is only known to occur between two successive visits. That is, all the observed data are interval-censored. Participants who do not experience conversion at their last follow-up date are right-censored. Also, it is known that even in this at risk population, some individuals will never experience conversion [6], therefore, a fraction of the population is "immune" to the event, or "cured", as opposed to "susceptible" or "uncured". It is interesting to identify which covariate impacts the probability of being susceptible or not, the time until the conversion, or both. We thus need a method that allows such variable selection and analysis. Up to now, these data have been analyzed without variable selection and without accounting for a possible cure fraction, but dealing with the interval-censored nature of the data.

Most statistical softwares propose methods for right-censored data, but few of them allow data to be interval-censored [7]. In a non-parametric setting, the Kaplan-Meier estimator is no longer available as, in most of the cases, the events can no longer be ordered. To overcome this, the Turnbull non-parametric survival estimator was elaborated [8], and only recently a generalization to allow for continuous covariates was proposed [9]. Regression models have also been studied under that type of censoring [10, 11, 12, 13, 14, 15]. However, all these methods usually make use of complex algorithms or methods, such as Expectation-Maximization (EM) algorithm [16] , self-consistency algorithm [8], Iterative Convex Minorant algorithm[12], or B-spline smoothing techniques [13]. On the contrary, assuming a specific distribution for the event times makes the analysis much simpler in the presence of interval-censoring.

When a fraction of the population is not susceptible, the survival distribution is improper, leading the survival function to level off at a value different from zero. In this case, estimation of the proportion of immune individuals is of primary importance. In the past decades, numerous authors have proposed alternatives to standard survival techniques to take a cure fraction into account. Pioneers in that field were [17] and [18]. They supposed the global population could be seen as a mixture of cured and susceptible individuals, leading to the mixture cure model. An alternative is the promotion time model [19, 20], which assumes an upper bound for the cumulative hazard, and hence is also called the bounded cumulative hazard model. It was developed to maintain the assumption of proportional hazards, and is based on a biological interpretation. In a mixture cure model, the incidence, i.e. the cure probability, is often modeled parametrically, usually via a logistic regression model, or more rarely via a logit or a probit model. Only very few attempts to model this part of the model non-parametrically have been proposed. Regarding the latency part, i.e. modeling the impact of covariates on the time to event of susceptible individuals, both parametric and semi-parametric models have been proposed. Semi-parametric models do not specify any distribution function [21, 22, 23, 24, 25]. These models, however, have the disadvantage to rely on the time-consuming EM algorithm for inference. Therefore, fully parametric mixture cure models, in which the latency is often modeled via a Cox PH model, in which the baseline hazard is defined parametrically [26], can be a good alternative. Another choice

for the latency part can be the accelerated failure time (AFT) model, for example when the hypothesis of proportional hazards is not met [27]. Besides, as Sir David Cox stated [28], "accelerated life models are in many ways more appealing because of their quite direct physical interpretation". In a parametric AFT model, a specific distribution is assumed for the error-term. To avoid strong assumptions with regard to this specification, the extended generalized gamma (EGG) has been proposed as a flexible choice [29, 30]. This distribution includes the normal and Weibull distributions, both widely used in survival analysis.

The mixture cure model also allows a direct interpretation of the effect of covariates on the cure probability, and on the survival distribution for susceptible individuals, separately. Interestingly, these two sets of covariates may not necessarily be the same, and the number of potential covariates to be included in each component of the model can be large. Variable selection is thus needed so that the final model possesses good predictability and can easily be interpreted. Classical variable selection methods, like the well known best subset or stepwise selection, suffer from some serious drawbacks. For example, the computational load increases with the increasing number of variable in the model, and the process is discrete and non-stable, as it either enters or deletes a covariate from the model. Several other drawbacks are described by [31, 32]. On the contrary, shrinkage methods, such as the LASSO [33] and adaptive LASSO [34] are continuous processes: the general idea is to shrink some coefficients towards zero. This allows simultaneous variable selection and coefficient estimation. Moreover, newly proposed algorithms, such as the LARS algorithm [35], the coordinate descent [36] and the unified algorithm with quadratic approximation [31], allow to obtain results in an efficient way.

To the best of our knowledge, no work in the literature dealing with a cure fraction and interval-censoring uses such a variable selection approach. Dealing with right-censoring only, the adaptive LASSO procedure was extended to a Cox mixture cure model [37]. The authors use the fact that a mixture cure model, in which a Cox proportional hazard is assumed in the latency, can be estimated iteratively in two parts: the Cox model and the logistic regression. In this context, the use of existing adaptive LASSO procedure for the Cox model, and for the logistic regression in the incidence is straightforward. However, such a split in parametric models is not feasible, so that existing methods can not be applied directly. Therefore, we believe that the extension of the adaptive LASSO in this case can really be convenient if, for example, one wants to use a specific distribution.

In this paper, we account for a fraction of immune individuals in the global population by assuming a mixture cure model, allowing to distinguish effects of covariates on the probability to experience the event, and on the survival times for susceptibles. To cope for a possible departure of proportional hazards and to ease interpretation of the results, we assume an accelerated failure time regression model for the latency part. The extended generalized gamma distribution is used for the error term and the maximum likelihood function can be derived while taking interval-censoring into account. This distribution has the advantage of being very flexible while avoiding the use of the EM algorithm. At last but no least, we extend the adaptive LASSO procedure to our mixture cure model to perform a continuous variable selection for each component of the model.

The paper is divided as follow: in Section 2, we describe the model, as well as the estimation method. Section 3 presents our extension of the adaptive LASSO to the presence of a cure fraction. We investigate the finite sample properties of the method via a simulation study in

Section 4. Lastly, we present results of the application of the method to the aforementioned Alzheimer's disease database in Section 5, and we end with a conclusion. We also provide an appendix with more simulation results.

2 Model and estimation method

2.1 Extended Generalized Gamma AFT model for uncensored data

Consider n independent subjects, and let $T_1, ..., T_n$ represent their event times. We assume the following transformed location-scale model,

$$\log(T) = \mu(\boldsymbol{\beta}, \boldsymbol{X}) + \sigma\varepsilon.$$

The location μ is parametrically defined through parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)^T$ and a m-vector of covariates $\boldsymbol{X}$. As stated in [30], the scale σ can also depend on covariates, but we will assume a constant form for more simplicity. ε is an error term with probability density function f_ε, and survival distribution S_ε. Assuming that $\mu(\boldsymbol{\beta}, \boldsymbol{X}) = \boldsymbol{X}^T\boldsymbol{\beta}$ leads to the classical accelerated failure time (AFT) model:

$$\log(T) = \boldsymbol{X}^T\boldsymbol{\beta} + \sigma\varepsilon.$$

Making the assumption that the error term ε is independent of the covariates $\boldsymbol{X}$, the conditional survival distribution of $T = t$, $S(t|\boldsymbol{x})$, is given by:

$$S(t|\boldsymbol{X} = \boldsymbol{x}) = S_\varepsilon\left(v_{\boldsymbol{\beta},\sigma}(t, \boldsymbol{X})\right) \tag{1}$$

where $v_{\boldsymbol{\beta},\sigma}(t, \boldsymbol{X}) = \frac{\log(t) - \boldsymbol{X}^T\boldsymbol{\beta}}{\sigma}$. The probability density function and survival distributions of ε are given by, respectively:

$$f_\varepsilon(v; q) = \begin{cases} \frac{|q|}{\Gamma(q^{-2})}(q^{-2})^{q^{-2}} \exp(q^{-2}(qv - e^{qv})) & \text{if } q \neq 0 \\ \frac{1}{(2\pi)^{1/2}} \exp(-v^2/2) & \text{if } q = 0 \end{cases} \tag{2}$$

and

$$S_\varepsilon(v; q) = \begin{cases} 1 - I(q^{-2}e^{qv}, q^{-2}) & \text{if } q > 0 \\ I(q^{-2}e^{qv}, q^{-2}) & \text{if } q < 0 \\ \int_v^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx & \text{if } q = 0, \end{cases} \tag{3}$$

where $I(\cdot, k)$ is the incomplete gamma integral, that is $I(\cdot, k) = \frac{1}{\Gamma(k)} \int_0^\cdot x^{k-1}e^{-x}dx$ [38]. The resulting conditional distribution of T is called the extended generalized gamma distribution. It covers a wide class of distributions and is negatively skewed if $q > 0$ and positively skewed if $q < 0$. It includes, as special cases, extensively used distributions in survival analysis, i.e. the log normal distribution ($q = 0$), the Weibull distribution ($q = 1$), the inverse Weibull ($q = -1$). Originally, the EGG was introduced by [39]. It was later re-parameterized to avoid, amongst others, boundary problems for the normal distribution. For more information, we refer to [38] and [40].

2.2 Logistic EGG-AFT model with interval-censored data and a cure fraction

In the presence of interval censoring, we do not observe $t_1, ..., t_n$. Rather, we observe l_i and r_i such that $t_i \in [l_i, r_i[$ for $i = 1, ..., n$. Note that right-censored observations are also covered if we allow $r_i = \infty$. We also assume independent censoring, conditional on the covariates. The contribution to the likelihood of each observation is $S(l_i) - S(r_i)$ for an interval censored observation and $S(l_i)$ for a right-censored one. We define the censoring indicator to be δ_i, with $\delta_i = 1$ if the observation i is interval-censored and $\delta_i = 0$ if it is right-censored.
In the mixture cure model, we assume that the population is a mixture of susceptible and cured individuals and we model separately the probability of being susceptible (the incidence) and the time-to-event for the susceptibles (the latency). First, denote by Y the variable such that $y_i = 1$ if individual i will get the event (susceptible) and 0 otherwise (cured). Due to censoring, the variable Y is only partially observed. The conditional probability to get the event is modeled by a logistic regression:

$$p(\boldsymbol{z}) = \mathbb{P}(Y = 1 | \boldsymbol{Z} = \boldsymbol{z}) = \frac{\exp(\boldsymbol{z}^T \boldsymbol{\gamma})}{1 + \exp(\boldsymbol{z}^T \boldsymbol{\gamma})},$$

where $\boldsymbol{Z}$ is a s-vector of covariates, not necessarily the same as those of $\boldsymbol{X}$, and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, ..., \gamma_s)^T$ is the corresponding vector of coefficients.
Second, the time-to-event for a susceptible individual is modeled with the EGG-AFT model. Denote by $S_u(\cdot|\boldsymbol{x})$ the survival distribution for the uncured individuals, given by (1) and (3). The conditional survival distribution for the global population is given by

$$S_G(t|\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z}) S_u(t|\boldsymbol{x}) + 1 - p(\boldsymbol{z}).$$

All interval censored observations are susceptible, and this occurs with probability p, their contribution to the likelihood is therefore $p(\boldsymbol{z})(S_u(l_i|\boldsymbol{x}) - S_u(r_i|\boldsymbol{x}))$. On the other hand, right-censored observations are either susceptible (with probability p), or actually cured (with probability $1 - p$); their contribution to the likelihood is then $p(\boldsymbol{z}) S_u(l_i|\boldsymbol{x}) + (1 - p(\boldsymbol{z}))$.
Writing $\boldsymbol{\eta} = (q, \boldsymbol{\beta}^T, \sigma, \boldsymbol{\gamma}^T)^T$, the log-likelihood function of the model is given by:

$$l_n(\boldsymbol{\eta}) = \sum_{i=1}^{n} \delta_i \left[\log(p(\boldsymbol{z_i})(S_u(l_i|\boldsymbol{x_i}) - S_u(r_i|\boldsymbol{x_i})))\right] + (1 - \delta_i) \left[\log(p(\boldsymbol{z_i}) S_u(l_i|\boldsymbol{x_i}) + (1 - p(\boldsymbol{z_i})))\right].$$

The likelihood function can be maximized using standard methods (e.g. Newton-Raphson) to obtain maximum likelihood estimates (MLEs) $\hat{\boldsymbol{\eta}} = (\hat{q}, \hat{\boldsymbol{\beta}}^T, \hat{\sigma}, \hat{\boldsymbol{\gamma}}^T)^T$. Theoretical large-sample properties of MLE's follow, such as consistency and unbiasedness. Also, the Hessian matrix provides an estimate of the variance-covariance matrix of $\hat{\boldsymbol{\eta}}$. Inference for latency and incidence parts is straightforward. In particular, a likelihood ratio test can be used to detect departure from a particular distribution included in the EGG, for example the Weibull or the log normal distributions [38, 41]. This way, a simpler model can always be reached when appropriate. For tests of the form $H_0 : q = q_0$ versus $H_1 : q \neq q_0$, the likelihood ratio statistic is

$$\Lambda = 2(l_n(\hat{\boldsymbol{\eta}}_0) - l_n(\hat{\boldsymbol{\eta}})),$$

where $\hat{\boldsymbol{\eta}}_0$ is the MLE assuming $q = q_0$. For finite q, the distribution of Λ under the null hypothesis asymptotically follows a chi-square distribution with one degree of freedom.

3 Variable Selection

3.1 The adaptive LASSO

Consider first the case of no cure fraction, that is, a simple EGG-AFT model with parameter $\boldsymbol{\eta} = (q, \boldsymbol{\beta}^T, \sigma)^T$. In this setting, penalized regression methods have been widely used and are based on a penalized log-likelihood of the form:

$$-l_n(\boldsymbol{\eta}) + n\lambda \sum_{j=1}^{m} p_j(|\beta_j|), \tag{4}$$

where $l_n(\boldsymbol{\eta})$ is the log-likelihood function. In the second term of (4), λ represents the penalty term (the tuning parameter), controlling for the amount of shrinkage of the estimates. If it is equal to zero, then minimizing (4) leads to the usual unpenalized MLE; otherwise, the coefficients are shrunk towards zero. The function $p_j(|\cdot|)$ is the penalty function and can take several forms (for example, the LASSO penalty [33], SCAD penalty [31], ridge penalty [42]). The adaptive LASSO penalty [34] is given by:

$$p_j(|\beta_j|) = |\beta_j| w_j,$$

with $\boldsymbol{w} = (w_1, \ldots, w_m)^T$ being a known weight vector. The adaptive LASSO is, as the LASSO, a convex optimization problem with l_1-norm, and any algorithm used to solve a LASSO problem can be easily adapted to the adaptive LASSO case [34], for example, the LARS algorithm [35]. Unlike the LASSO, the adaptive LASSO possesses the oracle property, as long as the weights w_j are data-dependent and cleverly chosen [34]. We follow the proposal of [43] to take $w_j = 1/|\hat{\beta}_j|$, where $\hat{\beta}_j$ is the unpenalized MLE, reflecting somehow the importance of corresponding covariates. Of course, any other consistent estimator can be chosen for $\hat{\beta}_j$, see [34] for guidance when, for example, there is collinearity issues.
The LARS algorithm was originally aimed at solving penalized least square problems. Nevertheless, any likelihood function can be expressed in an asymptotic least square equivalent, so that use of LARS algorithm is possible. Following [44], using Taylor expansion, $l_n(\boldsymbol{\eta})$ can be approximated by

$$l_n(\hat{\boldsymbol{\eta}}) + \frac{1}{2}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})^T \ddot{l}_n(\hat{\boldsymbol{\eta}})(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}),$$

where $\hat{\boldsymbol{\eta}}$ is an unpenalized consistent estimator, and $\ddot{l}_n(\hat{\boldsymbol{\eta}})$ represents the matrix of second derivatives of the log-likelihood at $\hat{\boldsymbol{\eta}}$. The following equation is the *least square approximation* (LSA) of the log-likelihood $l_n(\boldsymbol{\eta})$:

$$Q(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}) = (\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})^T \ddot{l}_n(\hat{\boldsymbol{\eta}})(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}). \tag{5}$$

The minimizer of -$Q(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}})$ is different from the estimates obtained by minimizing the minus log-likelihood, henceforth, the maximizer of (5) is called the LSA estimator [44].

3.2 The adaptive LASSO in the presence of cure individuals

In the presence of cured individuals $\boldsymbol{\eta} = (q, \boldsymbol{\beta}^T, \sigma, \boldsymbol{\gamma}^T)^T$ and the variables impacting the probability of being cured may not necessarily be the same as those impacting the survival

distribution of the susceptible people. Therefore, we propose to penalize both the incidence and the latency part, allowing a different penalty term in each part. This leads to the following minimization criterion:

$$-Q(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}) + n\lambda_\beta \sum_{j=1}^{m} \frac{|\beta_j|}{|\hat{\beta}_j|} + n\lambda_\gamma \sum_{j=1}^{s} \frac{|\gamma_j|}{|\hat{\gamma}_j|},$$

where s is the number of variables in the incidence part, λ_β is the tuning parameter for the β's, and λ_γ is the tuning parameter for the γ's.
To solve this optimization problem with the LSA estimator and the LARS algorithm, one can proceed iteratively in several steps. We optimize first with respect to the β's, holding every other parameter fixed, then do the same for the γ's. This way, we can easily obtain adaptive LASSO solutions, with two different penalty terms. We have the following algorithm:

Step 1. Obtain the unpenalized MLE $\hat{\boldsymbol{\eta}} = (\hat{q}, \hat{\boldsymbol{\beta}}^T, \hat{\sigma}, \hat{\boldsymbol{\gamma}}^T)^T$ by maximizing $l(\boldsymbol{\eta})$.

Step 2. Set $\boldsymbol{\eta} = (\hat{q}, \boldsymbol{\beta}^T, \hat{\sigma}, \hat{\boldsymbol{\gamma}}^T)^T$, i.e., every other parameters than $\boldsymbol{\beta}$ are fixed. Minimize $-Q(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}) + n\lambda_\beta \sum_{j=1}^{m} \frac{|\beta_j|}{|\hat{\beta}_j|}$ to get adaptive LASSO estimate $\tilde{\boldsymbol{\beta}}$.

Step 3. Set $\boldsymbol{\eta} = (\hat{q}, \tilde{\boldsymbol{\beta}}^T, \hat{\sigma}, \boldsymbol{\gamma}^T)^T$, i.e., every other parameters than $\boldsymbol{\gamma}$ are fixed. Minimize $-Q(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}) + n\lambda_\gamma \sum_{j=1}^{m} \frac{|\gamma_j|}{|\hat{\gamma}_j|}$ to get adaptive LASSO estimate $\tilde{\boldsymbol{\gamma}}$.

Step 4. Set $\boldsymbol{\eta} = (q, \tilde{\boldsymbol{\beta}}^T, \sigma, \tilde{\boldsymbol{\gamma}}^T)^T$, i.e., every other parameters than q and σ are fixed. Maximize the unpenalized likelihood $l(\boldsymbol{\eta})$ with respect to q and σ. We then have $\tilde{\boldsymbol{\eta}} = (\hat{q}, \tilde{\boldsymbol{\beta}}, \hat{\sigma}, \tilde{\boldsymbol{\gamma}})$.

Step 5. Repeat step 2 to 4 until convergence.

3.3 Tuning parameter selection and variance estimation

The choice of the optimal penalty $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_\beta, \hat{\lambda}_\gamma)$ is of crucial importance and is done via a BIC selection criterion [44]. First, for fixed λ_β and λ_γ, let $\tilde{\boldsymbol{\beta}}_{\lambda_\beta}$ and $\tilde{\boldsymbol{\gamma}}_{\lambda_\gamma}$ be the adaptive LASSO estimates with λ_β and λ_γ, respectively. We minimize

$$BIC(\boldsymbol{\lambda}) = -Q(\tilde{\boldsymbol{\eta}}_{\boldsymbol{\lambda}}, \hat{\boldsymbol{\eta}}) + \log(n) df_\lambda,$$

where $\tilde{\boldsymbol{\eta}}_{\boldsymbol{\lambda}} = (\hat{q}, \tilde{\boldsymbol{\beta}}_{\lambda_\beta}^T, \hat{\sigma}, \tilde{\boldsymbol{\gamma}}_{\lambda_\gamma}^T)^T$ and df_λ is the number of non-zero coefficients in $\tilde{\boldsymbol{\eta}}_{\boldsymbol{\lambda}}$. We then take

$$\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_\beta, \hat{\lambda}_\gamma) = \arg \min_{(\lambda_\beta, \lambda_\gamma)} BIC((\lambda_\beta, \lambda_\gamma)).$$

The minimization can be done via a grid search amongst selected values of λ_β and λ_γ and we take the combination leading to the smallest BIC. This procedure allows $\hat{\lambda}_\gamma$ to be different from $\hat{\lambda}_\beta$, therefore a different amount of shrinkage in the latency part and in the incidence part can be reached.
Standard errors for adaptive LASSO estimates are calculated based on a ridge regression approximation and on the sandwich formula for computing the covariance matrix of the estimates [33, 31, 34].

Denote H the matrix of second derivatives of the log-likelihood at $\tilde{\boldsymbol{\eta}} = (\hat{q}, \tilde{\boldsymbol{\beta}}, \hat{\sigma}, \tilde{\boldsymbol{\gamma}})$. Define

$$A = diag\left(1, 1, \frac{\lambda_\beta}{\tilde{\beta}_1^2}, \cdots, \frac{\lambda_\beta}{\tilde{\beta}_m^2}, 1, 1, \frac{\lambda_\gamma}{\tilde{\gamma}_1^2}, \cdots, \frac{\lambda_\gamma}{\tilde{\gamma}_s^2}\right).$$

Also, define

$$D = diag\left(1, 1, \frac{\mathbb{1}(\tilde{\beta}_1 \neq 0)\lambda_\beta}{\tilde{\beta}_1^2}, \cdots, \frac{\mathbb{1}(\tilde{\beta}_m \neq 0)\lambda_\beta}{\tilde{\beta}_m^2}, 1, 1, \frac{\mathbb{1}(\tilde{\gamma}_1 \neq 0)\lambda_\gamma}{\tilde{\gamma}_1^2}, \cdots, \frac{\mathbb{1}(\tilde{\gamma}_1 \neq 0)\lambda_\gamma}{\tilde{\gamma}_s^2}\right).$$

Then the sandwich formula gives the following estimated covariance matrix:

$$cov(\hat{\boldsymbol{\eta}}) = (H + A)^{-1} (H + D) H^{-1} (H + D) (H + A)^{-1}.$$

The estimated variance of a coefficient set to zero is equal to zero. More details about this equation can be found in [45].

4 Simulation studies

The first objective of the simulation study is to investigate the behavior of our method, and to discuss the impact of the amount of cured and right-censored observations on the results. Secondly, we study the performance of the likelihood ratio test to detect whether the true underlying distribution is either log-normal or Weibull. Finally, we evaluate the adaptive LASSO procedure described above, both in term of estimation and variable selection. We use an adaptation of LSA R code from [44] to get estimates.

4.1 Simulations setting

Data are generated from the EGG-AFT mixture cure model. We consider 3 different sets of parameter value to reach 3 different levels of cure and right-censoring, see Table 1.
As stated in Section 2.1, the scale σ may depend on covariates as well. Here, we simply allow for one covariate. For all 3 scenarios, event times for susceptible individuals are generated to follow an EGG-AFT distribution with:

$$\begin{aligned} \log(T|\boldsymbol{X}) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \exp(\alpha_0 + \alpha_1 X_1)\varepsilon \\ &= 4.1 - 0.2X_1 + 0.5X_2 - 0.5X_3 + \exp(-2 + 0.5X_1)\varepsilon, \end{aligned} \tag{6}$$

where $X_1 \sim \text{Bern}(0.5)$, $X_2 \sim \mathcal{N}(0, 0.16)$ and $X_3 \sim \mathcal{N}(0, 0.25)$, and ε has probability density function (2).
For the incidence part, the cure variable $Y|\boldsymbol{Z} \sim \text{Bern}(p(\boldsymbol{Z}))$ and

$$p(\boldsymbol{Z}) = \frac{\exp(\gamma_0 + \gamma_1 Z_1 + 0.5Z_2 - 0.5Z_3)}{1 + \exp(\gamma_0 + \gamma_1 Z_1 + 0.5Z_2 - 0.5Z_3)}, \tag{7}$$

with $Z_1 = X_1$, $Z_2 \sim \mathcal{N}(0, 2)$ and $Z_3 \sim \mathcal{N}(0, 0.25)$. Values for q, γ_0 and γ_1 are given in Table 1 for each scenario.
To simulate intervals in which T_i lies, $i = 1, \cdots, n$, we follow the idea of [30]. For each

i, generate $V_i \sim U[0, 25]$, the first visit. Also, fix a maximum number of visits, say K. Then, if $T_i < V_i$, set $L_i = 0$, $R_i = V_i$. Else, if $T_i > V_i + 4K$, the observation is right-censored; set $L_i = V_i + 4K$, $R_i = \infty$. Otherwise, there exists $k_i = 1, 2, 3, \cdots, K$ such that $V_i + 4(k_i - 1) \leq T_i < V_i + 4k_i$; in this case set $L_i = V_i + 4(k_i - 1)$ and $R_i = V_i + 4k_i$. For each scenario, the value of K is given in Table 1. We simulate 500 datasets of sizes $n = 200$, $n = 300$ and $n = 500$ for each scenario.

4.2 Simulations results

First, we analyze the datasets with our EGG-AFT mixture cure model, without considering any variable selection. For comparison purposes, we also analyze the data without considering a cure fraction with a simple EGG-AFT model. Tables 2, 3 and 4 show the results for $n = 200$, $n = 300$ and $n = 500$, respectively. For any sample size, the bias and MSE for the latency part, i.e. the $\hat{\beta}$'s, are low. However, for the smallest sample sizes ($n = 200$), the bias and MSE in the incidence part, i.e. the $\hat{\gamma}$'s, can be large, especially if the cure proportion is low compared to the right-censoring rate. Table 2 shows large bias for the first scenario, where the cure proportion is 20%, and the right-censoring rate is 40%. These bias and MSE are decreasing with the sample size. Obviously, we need enough information, that is, enough cured individuals, in order to discriminate between cured and susceptible, and thus, to be able to perform accurate estimation in the incidence part. Globally, for a fixed right-censored proportion, if the cure fraction increases, the MSE in incidence decreases. The opposite for a fixed cure proportion: the more the right-censoring, the higher the MSE.
Regarding likelihood ratio tests, the first null hypothesis is $H_0 \equiv q = 0$, i.e. the survival time of the susceptibles follow a log normal distribution; and the second one is $H_0 \equiv q = 1$, i.e. the survival time of the susceptibles follow a Weibull distribution. The α level of the test is fixed to 5%. In all cases, we report the proportion of times the null hypotheses are rejected. This is the observed power (level) of the test when H_1 (H_0) is true. It can be seen that in all cases, when H_0 is true, the observed level is close to 5%. When the true parameter q is equal to 0.5, i.e. in between the log-normal and the Weibull distribution, the observed power is less than 50% for small sample sizes, revealing the difficulty to discriminate between these distributions. But as the sample size increases, this power increases towards 100%, showing strong evidence against any two of these distributions.
Concerning the analysis with an EGG-AFT model when no cure fraction is taken into account (lower part of Tables 2,3 and 4), the bias are larger than when using the EGG-AFT mixture cure model, especially for parameters q and α. More results about the impact of cure and right-censoring proportion can be found in the appendix.

4.3 Simulation results: variable selection method

Finally, we assess the performance of the adaptive LASSO pertaining to variable selection and estimation. We simulated data as described in Section 4.1, and we added 10 standard normal variables in both latency and incidence parts, whose coefficients are truly zero. Tables 5, 6 and 7 show the results for $n = 200$, $n = 300$, and $n = 500$ for all 3 scenarios. The upper part shows bias and MSE for the truly non-zero coefficients, and the lower part gives the average

number of correct (resp., incorrect) zero's, i.e. the average number of times the adaptive LASSO sets a coefficient to zero when it truly is zero (resp., non-zero). In the simulations, the optimal tuning parameter $\boldsymbol{\lambda}$ was chosen via the BIC-type selection criterion from Section 3.3. Globally, those results reflect the same trend as the previous analysis, i.e. low bias and MSE except for small sample size ($n = 200$); and increasing bias and MSE when, for a fixed right-censored proportion, the cure proportion decreases.
Compared to the analysis without variable selection, for non-zero coefficients, we detect larger bias and MSE in incidence. Indeed, the coefficients are shrunk to zero and this implies that the estimates are biased.
For $n = 500$, we see that our method performs well for both coefficient estimation and variable selection. The average number of correct zero is very close to the optimal value of 10, in both latency and incidence parts. The average number of incorrect zero is very close to the optimal value of 0 in the latency part, and higher in the incidence part. This is explained by the fact that, in the logistic regression (7), some covariates (here, Z_2 and Z_3) do not have an impact on the cure probability. So, the adaptive LASSO procedure interestingly sets these coefficients to zero. As a consequence, the bias for $\hat{\gamma}_2$ and $\hat{\gamma}_3$ is slightly larger. The effect of cured proportion and right-censoring rate, concerning variable selection, follows the same trend as analyzed before: the number of correct zero slightly decreases when there is more right-censoring. Overall, the adaptive LASSO performs satisfactorily for estimation as well as for variable selection, as it includes variables that truly have an impact on the model.

5 Application on real data : Oxford Project To Investigate Memory and Aging (OPTIMA)

We apply our approach to the data of a study linked to Alzheimer disease [5]. The study aims at predicting the time until amnestic Mild Cognitive Impairment (aMCI, a possible precursor of Alzheimer disease) occurs, based on initial cognitive scores. There were 241 cognitively healthy patients included in the study, of which 91 converted (37.8%), and the other 150 (62.2%) were right-censored. At their first visit, all patients were given a test, the Cambridge Cognitive Examination (CAMCOG). CAMCOG measures the level of cognitive impairment and assesses MMSE (Mini Mental Stage Examination), orientation, comprehension, expression, recent memory, remote memory, learning, abstract thinking, perception, praxis, attention and calculation. Other covariates can be taken into account, such as the age, the years of total education, the gender, and the presence or absence of Apolipoprotein E4 (ApoE4), a gene known to increase the risk to develop Alzheimer disease [46].
Conversion to aMCI was determined by a neuropsychologist (see [5] for more details) at each visit, which took place in average every year and a half. The data were clearly interval-censored since conversion actually occurred between visits, and the exact date was not known. Considering interval-censoring only, these data were previously analyzed by [5], using a different approach. They found 3 significant variables. Two of them with a positive impact on time to aMCI-conversion: expression and learning scores at first visit, and one with a negative impact: the age at first visit. However, they did not use a specific model to acknowledge that a proportion of the patients will never convert to aMCI. This is why we propose to analyze the data with our method, which considers both interval censoring and a cure proportion.

Figure 1 shows the Turnbull [8] nonparametric survival estimator, taking interval-censoring into account. The curve shows a plateau with only one event after more or less 12.5 years, revealing the possibility that a fraction of the population would never have experienced the event.
In our analysis, 12 potential prognostic factors were included in the model, both in the latency and in the incidence part : MMSE, expression, remote memory, learning, attention, praxis, abstract thinking, perception, ApoE4 status, gender, age, and years of total education, resulting in a total of 26 parameters. We used the EGG-AFT cure mixture model to get unpenalized maximum likelihood estimates, and the adaptive LASSO procedure described in Section 3 to perform variable selection.
Table 8 shows the adaptive LASSO estimate, the standard error estimated using formula from Section (3.3), and the exponentiated estimates. This allows a direct interpretation of the impact of covariates, in terms of acceleration or deceleration of the time to the event in latency; and in terms of increase or decrease in odds for the incidence.
Focusing on susceptible people (the latency part), there are 3 variables increasing the expected duration, thus having a positive impact on the survival, by at least 15%: expression (38%), perception (20%) and education (16%). On the other hand, only the age shortens the duration by at least 15%: when age increases by 5 years, the expected time until conversion is shorten by 27%. For comparison, without considering cure, the covariate perception was not significant, whereas the learning variable was significant, with a positive impact on the survival. However, we see that learning still has a positive impact, but in the incidence part, reducing the risk to be susceptible. Three other variables have a positive impact on the probability to be susceptible: MMSE (-89%), praxis (-73%), ApoE4 Status (-43%). At the opposite, the abstract thinking (62%) and the total years of education (883%) have here a highly negative impact and significantly increases the odds ratio.
With these results, we estimated the average cure proportion in the whole sample to 20%. Analyzing these data taking a cure fraction into account lead to more information: first, the positive impact of the learning variable is now due to the fact that it reduces the probability to convert to aMCI. Second, we now consider other variables that have an impact: those impacting the probability to experience the disease.

6 Conclusion and Discussion

In this article, we consider the accelerated failure time model in a context where data are interval censored and where a fraction of the population is cured from the event of interest. In survival analysis, the Cox proportional hazards model is widely used, provided that the proportional hazards assumption is met. Typically, in these cases, survival curves do not cross with each other. In the presence of a cure fraction, even if the survival distribution for susceptibles truly comes from a PH model, curves can cross with each other [22]. To our knowledge, there is no method to distinguish crossing hazards that are due to the presence of cure from crossing hazards that are due to a true non proportionality in the latency. Using an AFT model circumvents this issue in addition of providing a straightforward interpretation of the results.
Parametric models are often criticized because a departure from the true underlying distribution can have substantial consequences. Nonetheless, in the presence of interval-censoring

and cure, it is very difficult to develop simple yet efficient estimation procedures without imposing parametric restrictions especially for high dimensional data. This is why a flexible distribution, capable of capturing a lot of characteristics, is an excellent compromise in this context.
Although widely used in the context of dimension reduction, when the number of covariates exceeds the number of observations, shrinkage methods are also useful in our context. Indeed, the number of covariates may be large, as a set of covariates can be included twice, i.e. in both parts of the model. This is why we believe that such shrinkage methods should be extended to the mixture cure model.
Different aspects were highlighted from the simulation studies. First, using a mixture cure model, when a cure fraction is truly present, reduces the bias in the latency part. Second, if sample size is small, and if there is not enough cured individuals compared with the right-censoring proportion, then the bias and MSE in the incidence part can be large. Thus, there is a trade-off between the gain in bias in the latency, and the instability of estimates in the incidence. It is clear that if not enough cured individuals are present in the database, the model will not be able to discriminate between the susceptible and cured ones. Also, making use of the mixture model results in a different interpretation. Covariates can have an impact on the survival, on the cure probability, or on the both. This lead to even more information about the event of interest.
In conclusion, our model and variable selection procedure offers flexibility as well as an easy way to interpret the results. Even more flexibility can be reached, and other variable selection procedures deserve more attention in parametric cure mixture models. Those are subject to future work.

Acknowledgment

The first three authors acknowledges financial support from the IAP research network P7/06 of the Belgian Government (Belgian Science Policy), and from the contract "Projet d'Actions de Recherche Concertées" (ARC) 11/16-039 of the "Communauté française de Belgique", granted by the "Académie Universitaire Louvain".

References

[1] Visser PJ, Verhey FRJ. Mild cognitive impairment as predictor for alzheimer's disease in clinical practice: effect of age and diagnostic criteria. *Psychological Medicine* 1 2008; **38**:113–122, doi:10.1017/S0033291707000554. URL http://journals.cambridge.org/article_S0033291707000554.

[2] Collie A, Maruff P, Shafiq-Antonacci R, Smith M, Hallup M, Schofield PR, Masters CL, Currie J. Memory decline in healthy older people: Implications for identifying mild cognitive impairment. *Neurology* 2001; **56**(11):1533–1538.

[3] De Jager C, Blackwell AD, Budge MM, Sahakian BJ. Predicting cognitive decline in healthy older adults. *American Journal of Geriatric Psychiatry* 2005; **13**(8):735–740, doi:10.1176/appi.ajgp.13.8.735.

[4] Weaver Cargin J, Collie A, Masters C, Maruff P. The nature of cognitive complaints in healthy older adults with and without objective memory decline. *Journal of Clinical and Experimental Neuropsychology* 2008; **30**(2):245–257, doi:10.1080/13803390701377829.

[5] Oulhaj A, Wilcock GK, Smith aD, de Jager Ca. Predicting the time of conversion to MCI in the elderly: role of verbal expression and learning. *Neurology* Nov 2009; **73**(18):1436–42, doi:10.1212/WNL.0b013e3181c0665f.

[6] Petersen RC, Doody R, Kurz A, Mohs RC, Morris JC, Rabins PV, Ritchie K, Rossor M, Thal L, Winblad B. Current concepts in mild cognitive impairment. *Archives of Neurology* 2001; **58**(12):1985–1992, doi:10.1001/archneur.58.12.1985.

[7] Gomez, Calle ML, Oller R, Langohr K. Tutorial on methods for interval-censored data and their implementation in R. *Statistical Modelling* 2009; **9**(4):259–297, doi:10.1177/1471082X0900900402.

[8] Turnbull BW. The Empirical Distribution Function with Arbitrarily Grouped , Censored and Truncated Data. *Journal of the Royal Statistical Society. Series B (Methodological)* 1976; **38**(3):290–295.

[9] Dehghan M, Duchesne T. A generalization of turnbull's estimator for nonparametric estimation of the conditional survival function with interval-censored data. *Lifetime Data Analysis* 2011; **17**(2):234–255, doi:10.1007/s10985-010-9174-9.

[10] Finkelstein DM, Wolfe RA. A Semiparametric Model for Regression Analysis of Failure Time Data. *Biometrics* 1985; **41**(4):933–945.

[11] Rabinowitz D, Tsiatis A, Aragon J. Regression with interval-censored data. *Biometrika* 1995; **82**(3):501–513, doi:10.1093/biomet/82.3.501. URL http://biomet.oxfordjournals.org/content/82/3/501.short.

[12] Pan W. Extending the iterative convex minorant algorithm to the cox model for interval-censored data. *Journal of Computational and Graphical Statistics* 1999; **8**(1):109–120. URL http://amstat.tandfonline.com/doi/abs/10.1080/10618600.1999.10474804.

[13] Komárek A, Lesaffre E, Hilton JF. Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics* 2005; **14**(3):726–745, doi:10.1198/106186005X63734. URL http://dx.doi.org/10.1198/106186005X63734.

[14] Sun J. *The Statistical Analysis of Interval-censored Failure Time Data*. Statistics for Biology and Health, Springer, New York, 2006.

[15] Chen D, Sun J, Peace K. *Interval-Censored Time-to-Event Data: Methods and Applications*. Chapman & Hall/CRC, London, 2012.

[16] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)* 1977; **39**(1):1–38. URL http://www.jstor.org/stable/2984875.

[17] Boag J. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)* 1949; **11**(1):15–53. URL http://www.jstor.org/stable/10.2307/2983694.

[18] Berkson J, Gage R. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association* 1952; **47**(259):501–515, doi:10.1080/01621459.1952.10501187. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10501187.

[19] Tsodikov A. A proportional hazards model taking account of long-term survivors. *Biometrics* 1998; **54**(4):1508–16, doi:10.2307/2533675. URL http://www.ncbi.nlm.nih.gov/pubmed/9883549.

[20] Tsodikov A, Ibrahim J, Yakovlev A. Estimating cure rates from survival data: An alternative to two-component mixture models. *Journal of the American Statistical Association* 2003; **98**:1063–1078, doi:10.1198/016214503000000001007.

[21] Taylor JMG. Semiparametric estimation in Failure Time Mixture Models Estimation. *Biometrics* 1995; **51**(3):899–907.

[22] Sy JP, Taylor JMG, Way DNA, Francisco SS. Estimation in a Cox Proportional Hazard Cure Model. *Biometrics* 2000; **56**(1):227–236, doi:http://dx.doi.org/10.1111/j.0006-341X.2000.00227.x.

[23] Peng Y, Dear KBG. A Nonparametric Mixture Model for Cure Rate Estimation. *Biometrics* 2000; **56**(1):237–243, doi:10.1111/j.0006-341X.2000.00237.x.

[24] Li CS, Taylor JMG. A semi-parametric accelerated failure time cure model. *Statistics in medicine* Nov 2002; **21**(21):3235–47, doi:10.1002/sim.1260. URL http://www.ncbi.nlm.nih.gov/pubmed/12375301.

[25] Zhang J, Peng Y. A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in Medicine* 2007; **26**:3157–3171, doi:10.1002/sim.2748.

[26] Farewell VT. The combined effect of breast cancer risk factors. *Cancer* 1977; **40**(2):931–6. URL http://www.ncbi.nlm.nih.gov/pubmed/890675.

[27] Wei L. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine* 1992; **11**:1971–1879, doi:10.1002/sim.4780111409. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.4780111409/abstract.

[28] Reid N. A conversation with sir david cox. *Statistical Science* 08 1994; **9**(3):439–455, doi:10.1214/ss/1177010394. URL http://dx.doi.org/10.1214/ss/1177010394.

[29] Yamaguchi K. Accelerated failure-time regression models with a regression model of surviving fraction: an application to the analysis of "permanent employment" in Japan. *Journal of the American Statistical Association* 1992; **87**(418):284–292, doi:DOI:10.

1080/01621459.1992.10475207. URL http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1992.10475207.

[30] Chen Ch, Tsay Y, Wu Y, Horng C. Logistic AFT location-scale mixture regression models with nonsusceptibility for left-truncated and general interval-censored data. *Statistics in medicine* 2013; **32**(24):4285–4305, doi:10.1002/sim.5845. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.5845/full.

[31] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 2001; **96**(456):1348–1360, doi:10.1198/016214501753382273. URL http://www.tandfonline.com/doi/abs/10.1198/016214501753382273.

[32] Harrell F. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Graduate Texts in Mathematics, Springer, 2001.

[33] Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B(Methodological)* 1996; **58**(1):267–288. URL http://www.jstor.org/stable/10.2307/2346178.

[34] Zou H. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* Dec 2006; **101**(476):1418–1429, doi:10.1198/016214506000000735.

[35] Efron B, Hastie T. Least angle regression. *The Annals of statistics* 2004; **32**(2):407 – 499, doi:10.1214/009053604000000067. URL http://projecteuclid.org/euclid.aos/1083178935.

[36] Friedman J, Hastie T. Pathwise coordinate optimization. *The Annals of Applied Statistics* 2007; **1**(2):302–332, doi:10.1214/07-AOAS131. URL http://projecteuclid.org/euclid.aoas/1196438020.

[37] Liu X, Peng Y, Tu D, Liang H. Variable selection in semiparametric cure models based on penalized likelihood, with application to breast cancer clinical trials. *Statistics in medicine* Oct 2012; **31**(24):2882–91, doi:10.1002/sim.5378. URL http://www.ncbi.nlm.nih.gov/pubmed/22733695.

[38] Lawless JF. Inference in the Generalized Gamma and Log Gamma Distributions. *Technometrics* 1980; **22**(3):409–419.

[39] Stacy E. A generalization of the gamma distribution. *The Annals of Mathematical Statistics* 1962; **33**(3):1187–192.

[40] Prentice RL. A log gamma model and its maximum likelihood estimation. *Biometrika* 1974; **61**(3):539–544.

[41] Peng Y, Dear KBG, Denham JW. A Generalized F Mixture Model for Cure Rate Estimation. *Statistics in Medicine* 1998; **17**:813–830.

[42] Hoerl AE, Kennard RW. Ridge regression: Applications to nonorthogonal problems. *Technometrics* 1970; **12**:69–82, doi:10.1080/00401706.1970.10488635.

[43] Zhang HH, Lu W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika* Aug 2007; **94**(3):691–703, doi:10.1093/biomet/asm037. URL http://biomet.oxfordjournals.org/cgi/doi/10.1093/biomet/asm037.

[44] Wang H, Leng C. Unified LASSO estimation by least squares approximation. *Journal of the American Statistical Association* 2007; **102**(479):1039–1048, doi:10.1198/016214507000000509. URL http://amstat.tandfonline.com/doi/full/10.1198/016214507000000509.

[45] Lu W, Zhang HH. Variable selection for linear transformation models via penalized marginal likelihood. *Institute of Statistics Mimeo Series 2580*, North Carolina State University 2006. URL http://www.stat.ncsu.edu/information/library/papers/tr2580.pdf.

[46] Corder E, Saunders A, Strittmatter W, Schmechel D, Gaskell P, Small G, Roses A, Haines J, Pericak-Vance M. Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer's disease in late onset families. *Science* 1993; **261**:921–923, doi:10.1126/science.8346443. URL http://www.sciencemag.org/content/261/5123/921.abstract.

Tables

Table 1: Parameter values for 3 levels of cure proportion and right-censoring

	Scenario 1	Scenario 2	Scenario 3
Cure proportion	20%	30%	40%
Right-Censoring	40%	40%	60%
q	0	0,5	1
γ_0	2	1	0,85
γ_1	-1	-0,2	-0,85
K	14	14	12

Table 2: Results of simulations for $n = 200$: Bias and MSE of the EGG-AFT mixture cure model in the upper part of the Table; rejection percentage of the likelihood ratio test in the middle; bias and MSE of the EGG-AFT model in the lower part.

Sample Size : n=200						
	(20% Cure, 40% RC)		(30% Cure, 40% RC)		(40% Cure, 60% RC)	
	Bias	MSE	Bias	MSE	Bias	MSE
	EGG-AFT Mixture Cure Model					
q	-0,091	0,185	0,013	0,137	0,044	0,325
β_0	-0,003	0,001	0,000	0,001	0,001	0,002
β_1	0,001	0,002	0,001	0,002	0,005	0,005
β_2	0,006	0,009	0,007	0,012	-0,000	0,026
β_3	-0,004	0,004	-0,003	0,007	-0,014	0,011
α_0	-0,041	0,022	-0,060	0,032	-0,084	0,088
α_1	0,025	0,033	0,012	0,028	0,017	0,060
γ_0	1,899	9,915	0,221	0,215	0,755	2,670
γ_1	-0,347	5,663	-0,073	0,341	-0,288	2,566
γ_2	0,239	0,147	0,040	0,017	0,088	0,026
γ_3	-1,397	5,027	-0,180	0,686	-0,684	1,341
	Likelihood Ratio Test					
True Value of q	q=0		q=0.5		q=1	
$H_0 \equiv q = 0$	8%		40%		57%	
$H_0 \equiv q = 1$	90%		35%		7%	
	EGG-AFT Model without Cure					
q	-1,182	1,535	-1,664	2,930	-1,601	2,897
β_0	-0,042	0,003	-0,028	0,002	0,005	0,004
β_1	0,037	0,003	-0,000	0,003	0,184	0,044
β_2	0,004	0,011	0,001	0,021	0,005	0,035
β_3	-0,001	0,006	-0,004	0,010	-0,007	0,016
α_0	0,139	0,038	0,492	0,262	0,677	0,490
α_1	0,208	0,072	0,037	0,036	0,210	0,098

Table 3: Results of simulations for $n = 300$: Bias and MSE of the EGG-AFT mixture cure model in the upper part of the Table; rejection percentage of the likelihood ratio test in the middle; bias and MSE of the EGG-AFT model in the lower part.

Sample Size : n=300						
	(20% Cure, 40% RC)		(30% Cure, 40% RC)		(40% Cure, 60% RC)	
	Bias	MSE	Bias	MSE	Bias	MSE
	EGG-AFT Mixture Cure Model					
q	-0,025	0,077	-0,014	0,074	0,030	0,165
β_0	-0,001	0,000	-0,001	0,001	0,000	0,001
β_1	-0,002	0,001	0,001	0,001	0,002	0,003
β_2	-0,001	0,006	0,001	0,009	0,002	0,017
β_3	-0,003	0,003	-0,003	0,004	-0,005	0,008
α_0	-0,032	0,011	-0,029	0,016	-0,052	0,047
α_1	0,009	0,019	0,016	0,019	0,017	0,034
γ_0	0,138	0,350	0,029	0,089	0,075	0,201
γ_1	-0,078	0,356	-0,001	0,125	-0,050	0,234
γ_2	0,018	0,013	0,013	0,008	0,016	0,009
γ_3	-0,048	0,695	-0,017	0,407	0,014	0,437
	Likelihood Ratio Test					
True Value of q	q=0		q=0.5		q=1	
$H_0 \equiv q = 0$	5%		47%		80%	
$H_0 \equiv q = 1$	98%		55%		6%	
	EGG-AFT Model without Cure					
q	-1,146	1,397	-1,629	2,761	-1,524	2,550
β_0	-0,041	0,002	-0,025	0,002	0,013	0,003
β_1	0,042	0,003	0,003	0,002	0,193	0,044
β_2	0,001	0,007	0,003	0,015	-0,004	0,026
β_3	-0,002	0,004	-0,001	0,006	-0,001	0,011
α_0	0,162	0,038	0,515	0,281	0,698	0,507
α_1	0,210	0,064	0,048	0,029	0,211	0,080

Table 4: Results of simulations for $n = 500$: Bias and MSE of the EGG-AFT mixture cure model in the upper part of the Table; rejection percentage of the likelihood ratio test in the middle; bias and MSE of the EGG-AFT model in the lower part.

Sample Size : n=500						
	(20% Cure, 40% RC)		(30% Cure, 40% RC)		(40% Cure, 60% RC)	
	Bias	MSE	Bias	MSE	Bias	MSE
	EGG-AFT Mixture Cure Model					
q	-0,027	0,056	-0,018	0,039	0,044	0,078
β_0	-0,001	0,000	-0,001	0,000	0,002	0,001
β_1	-0,000	0,001	-0,001	0,001	0,001	0,001
β_2	-0,001	0,004	-0,000	0,005	0,006	0,009
β_3	-0,002	0,002	0,001	0,002	-0,005	0,004
α_0	-0,018	0,007	-0,016	0,009	-0,038	0,025
α_1	0,007	0,010	0,010	0,011	0,010	0,017
γ_0	0,108	0,196	0,017	0,051	0,033	0,074
γ_1	-0,066	0,199	0,002	0,072	-0,028	0,099
γ_2	0,007	0,008	0,005	0,004	0,005	0,004
γ_3	-0,027	0,380	-0,034	0,246	-0,011	0,279
	Likelihood Ratio Test					
True Value of q	q=0		q=0.5		q=1	
$H_0 \equiv q = 0$	7%		71%		96%	
$H_0 \equiv q = 1$	100%		79%		6%	
	EGG-AFT Model without Cure					
q	-1,108	1,273	-1,573	2,536	-1,402	2,072
β_0	-0,039	0,002	-0,021	0,001	0,025	0,002
β_1	0,046	0,003	0,006	0,001	0,209	0,048
β_2	-0,004	0,005	0,003	0,008	-0,001	0,013
β_3	-0,002	0,002	0,001	0,004	-0,007	0,006
α_0	0,176	0,038	0,531	0,291	0,717	0,525
α_1	0,209	0,055	0,041	0,016	0,205	0,059

Table 5: Results of 500 simulations, with adaptive LASSO variable selection for $n = 200$.

Sample Size : n=200						
Param.	(20% Cure, 40% RC)		(30% Cure, 40% RC)		(40% Cure, 60% RC)	
	Bias	MSE	Bias	MSE	Bias	MSE
q	0,110	0,361	0,136	0,566	0,386	2,110
β_0	0,000	0,001	0,004	0,002	-0,008	0,003
β_1	0,017	0,003	0,017	0,004	0,042	0,011
β_2	-0,053	0,014	-0,061	0,024	-0,116	0,060
β_3	0,029	0,006	0,021	0,008	0,030	0,078
α_0	-0,169	0,066	-0,190	0,117	-0,341	0,428
α_1	0,067	0,047	0,080	0,068	0,125	0,153
γ_0	1,058	8,433	0,298	0,458	0,499	2,095
γ_1	-0,691	7,539	-0,079	0,306	-0,146	1,908
γ_2	-0,092	0,085	-0,127	0,034	-0,125	0,045
γ_3	0,204	1,865	0,347	0,407	0,366	0,248
	Average number of correct/incorrect zero's					
	Latency					
Correct	9,3		9,1		8,1	
Incorrect	0,0		0,0		0,2	
	Incidence					
Correct	9,2		9,6		9,7	
Incorrect	1,8		2,3		1,9	

Table 6: Results of 500 simulations, with adaptive LASSO variable selection for $n = 300$.

Sample Size : n=300						
Param.	(20% Cure, 40% RC)		(30% Cure, 40% RC)		(40% Cure, 60% RC)	
	Bias	MSE	Bias	MSE	Bias	MSE
q	-0,004	0,141	0,056	0,150	0,123	0,492
β_0	-0,001	0,001	0,000	0,001	0,002	0,002
β_1	0,012	0,002	0,011	0,002	0,026	0,006
β_2	-0,037	0,008	-0,038	0,014	-0,056	0,024
β_3	0,015	0,003	0,017	0,005	0,016	0,009
α_0	-0,083	0,020	-0,112	0,040	-0,160	0,126
α_1	0,038	0,026	0,051	0,028	0,078	0,061
γ_0	0,551	2,231	0,116	0,151	0,272	0,455
γ_1	-0,274	1,119	0,025	0,109	-0,023	0,469
γ_2	-0,082	0,064	-0,111	0,026	-0,101	0,027
γ_3	0,385	0,397	0,366	0,175	0,348	0,198
	Average number of correct/incorrect zero's					
	Latency					
Correct	9,7		9,5		9,2	
Incorrect	0,0		0,0		0,1	
	Incidence					
Correct	9,6		9,9		9,8	
Incorrect	1,7		2,3		1,7	

Table 7: Results of 500 simulations, with adaptive LASSO variable selection for $n = 500$.

Sample Size : n=500						
Param.	(20% Cure, 40% RC)		(30% Cure, 40% RC)		(40% Cure, 60% RC)	
	Bias	MSE	Bias	MSE	Bias	MSE
q	-0,035	0,065	0,030	0,060	0,175	0,215
β_0	-0,002	0,000	0,000	0,000	0,005	0,001
β_1	0,005	0,001	0,007	0,001	0,017	0,003
β_2	-0,025	0,005	-0,032	0,007	-0,041	0,013
β_3	0,010	0,002	0,011	0,003	0,015	0,005
α_0	-0,038	0,010	-0,062	0,016	-0,135	0,062
α_1	0,020	0,015	0,028	0,013	0,039	0,025
γ_0	0,275	0,368	0,075	0,062	0,092	0,099
γ_1	-0,219	0,423	0,046	0,064	0,023	0,140
γ_2	-0,085	0,021	-0,060	0,013	-0,071	0,015
γ_3	0,353	0,203	0,330	0,196	0,360	0,166
	Average number of correct/incorrect zero's					
	Latency					
Correct	9,8		9,8		9,6	
Incorrect	0,0		0,0		0,0	
	Incidence					
Correct	9,8		9,9		9,8	
Incorrect	1,4		1,8		1,3	

Table 8: aMCI results: adaptive LASSO (aLASSO) estimates, standard errors and exponentiated estimates. Last column gives the increase in time-to-the-event (for the latency) and odds ratio (for incidence).

	Parameter	aLASSO	SD	Exp(Estimate)
Latency	Intercept (Lat.)	2,628	0,155	
	MMSE	-	-	
	Expression	0,321	0,077	1,38
	Remote	-	-	
	Learning	-	-	
	Attention	-0,057	0,043	0,94
	Praxis	-	-	
	Abstract Thinking	0,086	0,034	1,09
	Perception	0,182	0,052	1,20
	APOE E4	-0,092	0,025	0,91
	Gender	-0,061	0,022	0,94
	Age (5y.)	-0,321	0,144	0,73
	Total Education	0,152	0,163	1,16
Incidence	Intercept (Inc.)	2,657	0,985	
	MMSE	-2,250	0,802	0,11
	Expression	-	-	
	Remote	-	-	
	Learning	-0,969	0,425	0,38
	Attention	-	-	
	Praxis	-1,302	0,564	0,27
	Abstract Thinking	0,483	0,456	1,62
	Perception	-	-	
	APOE	-0,556	0,264	0,57
	Gender	-	-	
	Age (5y.)	-	-	
	Total Education	2,285	2,081	9,83

Figure Captions

Figure 1. Turnbull Survival Curve, taking interval-censoring into account